Psychon Bull Rev (2014) 21:332–343
DOI 10.3758/s13423-013-0508-y

THEORETICAL REVIEW

The conditional-recency dissociation is confounded with nominal recency: Should unitary models of memory still be devaluated?

Rani Moran · Yonatan Goshen-Gottstein

Published online: 13 September 2013

Abstract The conditional-recency dissociation between immediate and delayed free recall FR; Farrell (Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 324–347, 2010) has critical implications regarding the prolonged debate between unitary and dual-store models of memory. In immediate FR, when the availability of items is controlled for, the recency of the final list item increases across the first few output positions. No such increase is found in delayed FR, with a trend in the opposite direction. This dissociation challenges temporal context TCM; Howard & Kahana (Journal of Mathematical Psychology, 46, 269–299, 2002) and distinctiveness SIMPLE; Brown, Neath, & Chater (Psychological Review, 114, 539–576, 2007) unitary models of memory and suggests the involvement of a short-term buffer in immediate FR. We show that this dissociation is confounded with the different magnitudes of nominal recency (i.e., the prevalence of the final list item) found in immediate as compared to delayed FR. By reshuffling output orders and comparing the empirical results to those of a null hypothesis of no output-order effect, we controlled for the greater prevalence of the final list item that has been observed in immediate FR. Under this control, we found no evidence for a dissociation in the tendency to recall the final list item across output positions. This finding suggests that the conditional-recency dissociation imposes no new constraint on unitary models of memory. More generally, we demonstrate how biases that influence measures of output-order tendencies (e.g., conditional recency) can be controlled for, thus yielding "purer" measures of these variables.

Keywords Free recall · Recency · Output position · Contextual reinstatement · Temporal models of memory · Short-term buffer · Computational modeling · Permutation test

One of the longest debates in the history of memory research has been between unitary-store models, which do not postulate the existence of a short-term memory (STM) buffer, and dual-store models, which do postulate the existence of such a buffer. The current version of this debate was well reflected in a series of articles (Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Kahana, Sederberg, & Howard, 2008; Sederberg, Howard, & Kahana, 2008; Usher, Davelaar, Haarmann, & Goshen-Gottstein, 2008). The debate focused on two points. The first was theoretical, centering on the question of whether unitary models implicitly incorporate a short-term buffer or are truly unitary. The second point was empirical, relating to the abilities of the two types of models to adequately fit the existing body of data. Although the debate was in no way resolved, there is solace in the conformity of opinion within the articles regarding the critical data that need to be addressed—specifically, data that describe the multiplicity of findings regarding the recency effect. Indeed, both positions of the debate interpreted the very same set of key empirical phenomena, but in entirely different ways.

The present article addresses a novel empirical finding, the conditional-recency dissociation between immediate and delayed free recall (Farrell, 2010). This dissociation was revealed when the calculation of the probability of retrieval of the final list item as a function of output position took into account only those trials in which the final list item had not already been recalled in a previous output position; hence, the final list item was—so to speak—still "available" for retrieval

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0508-y) contains supplementary material, that is available to authorized users.

R. Moran (✉) · Y. Goshen-Gottstein
School of Psychological Sciences, Tel Aviv University, Ramat Aviv, POB 39040, Tel Aviv 69978, Israel
e-mail: rani.moran@gmail.com

Y. Goshen-Gottstein
e-mail: goshengott@gmail.com

(see the next section for a detailed numeric example and rationale). The computation of recency, conditional on the availability of the final list item for retrieval, is aptly dubbed "conditional recency." This dissociation was cited in the debate as providing evidence for a dual-store approach, and critically, it seems to be the only known recency-related effect that ostensibly cannot be interpreted by unitary models. Below, we describe this dissociation and the challenge that it poses to unitary models. It is important to note from the outset that Farrell conducted further analyses, ranging beyond this dissociation, to which we will return in the General Discussion.

In this article, we will show that the conditional-recency dissociation may be the result of a subtle confound with nominal recency. Moreover, once nominal recency is controlled for, the conditional-recency dissociation is no longer found.[1] On the positive side, our control method isolates conditional recency from the modulating effects of nominal recency. It thus provides both theoretical clarity and the prospect of benefiting other research that involves output-order effects.

Our analysis poses an interesting theoretical and methodological question. Farrell's (2010) investigation embodies a focal empirical measure (conditional recency) that unitary models fail to adequately fit, thereby compromising their status. This failure remains, even after our demonstration that this measure is confounded by nominal recency. Critically, once this confound is controlled for, the key dissociation—conditional recency—is no longer statistically significant. This raises the question of whether the status of extant theories and models should still be compromised, even though they fail to fit empirical effects that are nonsignificant once an adequate control is applied.

On the one hand, it could be argued that issues of confounds are important with respect to the theoretical validity of empirical effects, but are less so with respect to model fitting. Thus, a model's failure to fit any behavioral aspect of the data should count as evidence against its validity. This holds, even if this aspect confounds several psychological influences or if it describes an effect that, once an appropriate control is applied, does not reach an adequate level of statistical significance.

On the other hand, it is reasonable to assume that had the confound been uncovered "in time," then data fitting of the null effect would not have been undertaken in the first place, let alone have survived the peer-review process. We suggest that in the current state of research, any model is at best a "crude" approximation of actual mnemonic processes, and that currently no model can be expected to fit the entire plethora of empirical patterns. Thus, we posit that for the purpose of scientific clarity, it is constructive to place the hurdles of theoretical validity and statistical significance on empirical effects before devaluating models that fail to track them. If it is discovered that model-devaluating effects are indeed confounded with undesired influences, then the model's initial validity should be reinstated.

We find merit in both of these views. Forced to choose between them, we side with the latter. Nevertheless, we believe that the issue is not clear cut, and that it is important for the scientific community to address this question, in that it influences the way that science is conducted and published, and because it hinges on issues of the philosophy and sociology of science.

A tale of two recencies: Nominal recency and conditional recency

In immediate free recall (FR), participants study a list of items (e.g., words) and are then asked immediately to recall all items that they can, in any order. The robust finding is that items presented in the last serial positions of the study list—recency items—are recalled with higher probability than are pre-recency items: This is the (nominal) *recency effect* (Murdock, 1962). This finding is revealed by a monotonic increase in the "recency portion" of the serial-position curve.

In delayed FR, in contrast, the study phase is followed by a delay, during which an attention-demanding distractor task is given, so as to prevent participants from rehearsing the items. Critically, relative to immediate FR, recency effects in delayed FR are reduced or entirely eliminated (e.g., Glanzer & Cunitz, 1966; Postman & Phillips, 1965). This dissociation between immediate and delayed FR with respect to recency is still considered by some (Davelaar et al., 2005; Lehman & Malmberg, 2013) as being support for dual-store models of episodic memory, comprising a short-term and a long-term store. These models assume that recency items reap their gain in immediate FR by being retrieved from the "memory buffer" or the short-term store. In delayed FR, on the other hand, distractor activity presumably eliminates the possibility of rehearsing the items or replaces the item information in the buffer, and thus the advantage of recency items is reduced or even eliminated (Atkinson & Shiffrin, 1968).

However, this dissociation can also be interpreted by theories that conceptualize memory as a unitary entity. Specifically, as compared to pre-recency items, recency items may enjoy either enhanced overlap between their encoding context and the retrieval context (TCM; Howard & Kahana, 2002) or enhanced discriminability (SIMPLE; Brown, Neath, &

[1] Note that the second author of this article has, in the past, advocated the notion that the recency effect reflects the operation of an STM store (Davelaar et al., 2005; Usher et al., 2008; but see Howard, Venkatadass, Norman, & Kahana, 2007). Indeed, Farrell's (2010) novel dissociation and modeling of the data bolsters this very notion. Still, here we argue that this dissociation cannot tip the balance of evidence between unitary- and dual-store models and does not provide autonomous corroboration for the existence of a short-term store.

Chater, 2007). Such enhancements may mediate the facilitated recall of recency items in immediate FR and, in addition, may dissipate during the end-of-list distraction task, resulting in reduced recency in delayed FR (cf. the ratio rule; Bjork & Whitten, 1974; Glenberg et al., 1980).

In contrast to the standard (nominal) recency effect, the novel empirical dissociation discovered by Farrell (2010) is that of *conditional* recency. Farrell found that in immediate FR, the tendency to recall the final list item—that is, the one presented last in the study list—increases as output position increases. No such increase is found in delayed FR, and if anything, those data reveal a decreasing tendency to recall the last item as a function of output position.

Let us clarify some terminology. Conditional-recency is computed by controlling for the *availability* of the final list item. Throughout the article—and as defined by Farrell—when referring to the "availability" of the final list item we mean that the final list item had not been recalled in previous output positions in the current recall trial. Note that the term *trial* refers to the entire set of recalled items from a single studied list. Individual items recalled within the trial are designated by their corresponding output positions. To illustrate an imaginary recall trial, a participant may have studied an eight-word list (here each letter represents a unique word): A, B, C, D, E, F, G, and H. Later, at test, if she recalled five items in the following order—B, H, D, C, A—then these five items would constitute a single recall trial, with "B" appearing in Output Position 1, "H" in Output Position 2, and so forth. The term *serial position* refers to the order of the individual items at study, so in this example, "A" appears in Serial Position 1, "B" in Serial Position 2, and "H" is the final list item.

To illustrate the computation of conditional recency, consider a participant who studied 100 unique lists and performed 100 delayed FR trials of these lists. Suppose that she recalled the final list item in the Output Position 1 in 20 trials and in Output Position 2 in 10 trials. Because no output precedes the first recall, the final list item was available on all 100 trials. Therefore, for Output Position 1, the ensuing conditional (as well as nominal) probability is 20 %. Considering Output Position 2, the final list item was available only on the remaining 80 trials. Therefore, whereas the nominal probability of recalling the final list item is 10 % (10 out of 100 trials), the conditional probability is 12.5 % (10 of the remaining 80 trials).

In this example, the 20 % (conditional) probability of recall in Output Position 1 of the final list item was larger than recall of this item in Output Position 2 (12.5 %). This pattern mirrors that reported by Farrell (2010) in delayed FR, in which, conditional on availability, the probability of recall of the final list items showed a tendency to *decrease* as the output unfolded. Surprisingly, Farrell found a dissociation between immediate and delayed FR, with an effect in the opposite direction in immediate FR: an *increased* tendency to recall final list items as retrieval unfolded. To illustrate, had the final list item been recalled in Output Positions 1 and 2 in 60 and 30 (typical values for immediate FR)—not 20 and 10 (typical values for delayed FR)—of the 100 trials, respectively, the conditional probability in Output Position 1 would have been 60 %, and would have increased to 30 out of the 40 remaining trials, equaling 75 %, in Output Position 2. Together, these examples demonstrate a pattern qualitatively similar to Farrell's dissociation, of either a decreased (delayed FR) or an increased (immediate FR) tendency to recall the last item across output positions.

Unlike Farrell's (2010) finding of an increased tendency to recall final list items as retrieval unfolds in immediate FR, unitary theories of recall predict a decreased tendency. This is because as retrieval unfolds, so does the passage of time, thereby diminishing the overlap between the encoding context of the final list item and the evolving retrieval context, relative to prefinal list items (or rendering the final list item less discriminable).[2] Indeed, Farrell demonstrated that the dissociation challenges unitary models, which predict that the deeper into the test-phase, the lower the probability of recalling the final list item. In contrast, models that include a short-term buffer—which drives the first few outputs in immediate but not in delayed FR—can capture this dissociation.

The dissociation between immediate and delayed FR in the propensity to recall the final list item as retrieval unfolds, emerges when computing conditional, rather than nominal, probabilities. But what is the rationale for computing conditional rather than nominal recall probabilities? When studying the effects that passage of time during the recall phase exerts on the propensity to recall the final list item (in either immediate or delayed FR), nominal probabilities for recalling this item in different output position are inadequate. The reason for this inadequacy stems from the fact that participants are, to quote Farrell (2010), "unwilling to report prior recalls." In

[2] To be sure, this assertion is overly simplistic. Deriving predictions for these models is far from straightforward. Consider SIMPLE, for example (similar considerations are relevant with respect to TCM). For a specific trial, changes in the discriminability of the final list item as output unfolds depend on the realized recall history (i.e., which items have already been recalled) and on the specific model parameters (i.e., alternative parameter values could generate different predictions). Thus, definitive model predictions can be formally derived only once the model is fitted to empirical recall data (yielding estimates of the actual parameters), and simulations are conducted on the basis of estimated parameter values. Indeed, such an analysis was conducted by Farrell (2010). Throughout the present article, whenever we discuss model predictions with respect to conditional recency, we will rely on the results of Farrell's analysis. By suggesting reasons for these predictions (diminished discriminability or contextual overlap), we aim at capturing the gist of the operative mechanisms, hopefully providing readers with helpful intuitions. Again, by no means does this eliminate the necessity for a formal analysis.

other words, if the final list item had already been recalled, participants will attempt to avoid repeating it in subsequent output positions. Thus, nominal probabilities confound the focal effect of temporal delay, caused by the passage of time, with the reduced tendency to recall the final list item due to attempts to avoid a recall repetition. By extracting only those trials on which the final list item has not yet been recalled, conditional probabilities eliminate this confound, enabling a more appropriate estimate of the effects of temporal delay.

In the sections that follow, we describe the conditional-recency dissociation more precisely, and explain how this dissociation is mathematically modulated—confounded—by "overall recency"—that is, by the overall probability of recalling the final list item.

Details of the analysis of conditional recency

The dissociation between immediate and delayed FR in conditional recency (Farrell, 2010) is based on an analysis of 14 conditions taken from five experiments (Howard & Kahana, 1999; Howard, Venkatadass, Norman, & Kahana, 2007; Murdock, 1962; Murdock & Okada, 1970). For each participant (in each condition) and for each output position k, the recency recall probability (RRP) was calculated as the proportion of trials in which the final list item was recalled in output position k, given that it had not already been recalled in the first $k - 1$ output positions. RRP measures the tendency to recall the final list item in each output position, conditional on its availability. On each trial, Farrell considered all recalls up to the first error (i.e., prior-list or extralist intrusions or repetitions), up to a maximum of four recalls.[3] For the illustrative recall trial described above, in which our imaginary participant recalled B, H, D, C, A, Farrell would have considered only the following items: B, H, D, C, chopping from the analysis the fifth item, A. Next, using simple linear regression, the slope of the RRP function across output positions was estimated. At a descriptive level, a negative slope indicates the tendency of recency for the final list item to decrease across output positions, as predicted by unitary models. A positive slope, in contrast, indicates an increasing tendency (for further details, see Farrell, 2010). Figure 1 shows the striking results of this analysis, with the mean slope being described as the dependent variable. Examination of the figure reveals a strong tendency of recency for the final list item to increase across initial output positions in immediate FR, whereas in delayed FR the trend is toward a decreasing, albeit nonsignificant, recency for the final list item. This pattern was supported formally by the appropriate multilevel linear regression analysis.

Next, to mitigate concerns of confounds due to the diverse experimental methodologies across studies, Farrell undertook a detailed examination of the single experiment that had used a repeated measures design for both immediate and delayed recall (Howard & Kahana, 1999, Exp. 1). Here, too, a higher mean RRP slope was found in immediate than in delayed FR.

Farrell (2010) also noted that the magnitude of the RRP slope, which ostensibly reflects the operation of the mechanism underlying recall, may be positively biased by the mathematical property of the diminishing size of the pool of available items as output unfolds. For example, for a 20-word list, under total random recall, the conditional probability (and also the nominal probability) of recalling the final list item in Output Position 1 is 1/20, which increases to 1/19 for Output Position 2 simply because, if the final list item is still available, it is now one of 19, rather than 20, candidates for output.

To address this concern, Farrell (2010) calculated for each participant a control slope (see the dashed lines in Fig. 1) and subtracted these slopes from the RRP slopes. Examination of the figure reveals that the diminishing pool of available words biases RRP slopes only slightly and cannot account for the dissociation between immediate and delayed FR. This dissociation, therefore, must reflect an intrinsic property of the retrieval dynamics of free recall. We next argue that the diminishing word-pool size was not the only confound that required control.

Conditional recency is confounded with recency

The main purpose of the conditional-recency slope is to gauge the tendency of the final list item to be recalled in different output positions. However, as we now explain, it turns out that this slope is highly modulated by the level of nominal recency for the final list item and that this modulation is—like in the case of a diminishing word pool—a mathematical property of the measure rather than a function of the theoretical construct that conditional recency purports to index. We then present a novel method that estimates the tendency of the final list item to manifest as output unfolds, free from the mathematical biasing effects of nominal recency.

Consider a "hypothetical participant," reflecting an imaginary psychological mechanism, who recalls items in different serial positions according to a fixed probability distribution (i.e., the probabilities described by the classic serial-position curve). Critically, the participant recalls items in a random output order. Specifically, across trials, the final list item appears with equal probability in each of the output positions. For this participant—irrespective of the prevalence of the final list item—the final list item should show no special tendency to manifest in any particular output position, because output

[3] Following the fourth output position, some of the RRP curves changed direction (e.g., from increasing to decreasing; see, Farrell, 2010, for a detailed analysis).

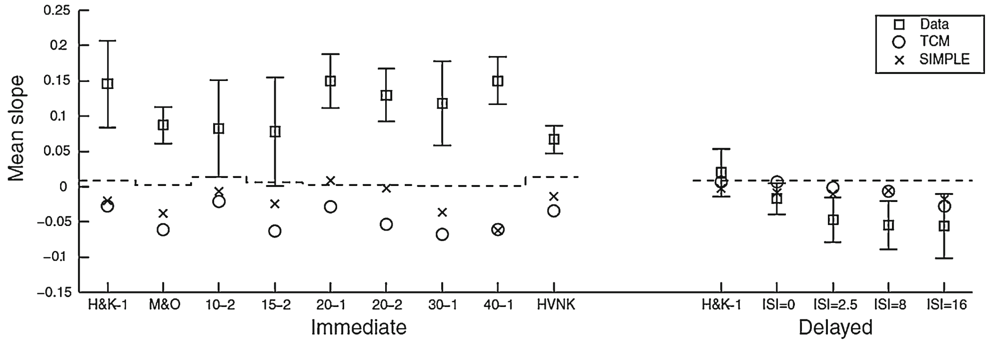

Fig. 1 Mean slopes predicting the conditional next-recall probability of the last list item from output position, for Output Positions 1–4. Immediate free recall conditions are grouped on the left, and delayed free recall conditions are grouped on the right. For each condition, results are displayed for the data (squares), the temporal context model (TCM; circles), and SIMPLE (crosses). The conditions are, from left to right, the immediate recall condition of Experiment 1 of Howard and Kahana (1999; H&K); the experimental condition of Murdock and Okada (1970; M&O); the six conditions of Murdock (1962), with the first number of each label giving list length, and the second, the presentation time per item, in seconds; the experimental condition of Howard et al. (2007; HVNK); the delayed free recall condition of Experiment 1 of Howard and Kahana (1999); and the four conditions of Experiment 2 of Howard and Kahana (1999), with the labels indicating the durations of continuous distraction between items at presentation. Error bars show single-sample confidence intervals. ISI, interstimulus interval (in seconds). From "Dissociating Conditional Recency in Immediate and Delayed Free Recall: A Challenge for Unitary Models of Recency," by S. Farrell, 2010, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, p. 326. Copyright 2010 by the American Psychological Association. Reprinted with permission

order is random. It turns out, however, that in contrast to the anticipated absence of such tendency, the RRP slope of our hypothetical participant does vary, and considerably so, as a function of the prevalence of the final list item (i.e., the overall nominal recency). Thus, an appropriate control for the prevalence of the final list item is required before conclusions from the pattern of the RRP-slope dissociation can be drawn.[4]

To demonstrate this idea, and for simplicity's sake, consider the case in which our participant recalls two items per trial. Let us first consider a case in which the nominal rate of recency for the final list item is 30 % (not unlike that found in delayed FR). Here, our participant would recall the final list items in Output Position 1 in 15 % of the trials, and in Output Position 2 in another 15 % of the trials, with Farrell's measure of conditional recency yielding the values of RRP(1) = .15, RRP(2) = .15/.85, and a miniscule slope—computed as RRP(2) – RRP(1)—equal to .0265. In contrast, if the nominal rate of recency for the final list item were 80 % (not unlike that found in immediate FR), with 40 % in each of the two output positions, the calculations would yield RRP(1) = .4, RRP(2) = .4/.6, with a slope more than 10 times larger (.2667) than that found under a nominal recency of 30 %.

Thus, the RRP slope varies considerably as a function of the frequency of the final list item. If so, it is possible that the same underlying (decreasing) tendency of the final list item to manifest across different output positions exists in both immediate and delayed FR and that the larger (positive) slope observed in immediate FR may simply reflect the amplified bias, due to the larger levels of overall recency for the final list item, in this condition.

Let us be clear. In arguing that recency and conditional recency are confounded, we mean that if the same propensity to recall the final list item across output positions is operative in both immediate and delayed FR, then a difference in the overall level of recency for the final list item would bias the slopes such that a higher empirically estimated slope for immediate FR would be found. Importantly, one must not misinterpret this *modulating* relation by assuming that conditional recency is simply an *equivalent* measure of the overall level of recency. Indeed, highly variable levels of conditional recency could be realized due to the intrinsic characteristics of retrieval dynamics, even when the overall level of recency for the final list item is invariant, and hence biases conditional recency to the same extent.[5]

[4] Note that Farrell (2010) controlled for the diminishing size of the pool of available items, but not for the prevalence of the final list item.

[5] Suppose, for example, that a participant always recalls two items and that the final list item is recalled in Output Position 1 on 10 % of the trials, and in Output Position 2 on 20 % of the remaining 90 % of the trials. In total, the final list item is recalled on 10 % + 18 % = 28 % of the trials, RRP(1) = .1, RRP(2) = .2, and the RRP slope is .1. On the other hand, if the final list item is never recalled in Output Position 1 but is recalled on 28 % of the trials in Output Position 2, then the overall level of recall (28 %) of the final list item is invariant, but now RRP(1) = 0, RRP(2) = .28, and the slope is .28.

Different models of memory may produce similar levels of overall recency and yet differ substantially in the conditional-recency slopes that they produce. See the simulations reported in Farrell (2010), in which both the TCM and the forward-buffer version did reasonable jobs of capturing overall recency, but differed substantially in the slopes that they produced.

A reanalysis of condition recency when controlling for nominal recency for the final list item

We now present a new control, the *random-order control*, which provides a general method of controlling for interfering influences when measuring any variable (e.g., the RRP slope) that embodies a theoretical tendency that manifests across output positions (we denote such a variable as V). After presenting the method, we will apply it to the conditional-recency dissociation.

The first step is to compute the variable V (e.g., the conditional-recency slope) from the empirical data in order to obtain the empirical measurement V_{data}. The empirical measure V_{data} might be influenced by confounding variables that do not reflect output order per se, and thus that might bias it (e.g., the last-item nominal level of recency, the total number of items recalled, or the length distribution of the recall trials). To control for these undesired influences, we randomly permute each of the recall trials of the participant. The permuted set of trials can be viewed as reflecting an imaginary participant who is equivalent to the actual participant with respect to all of the structural components of the recall sequence, save output order. We then calculate V for the participant in *the exact same manner* as V_{data} was calculated, but for the permuted instead of the empirical data—thus obtaining $V_{\text{perm}(1)}$. By repeating this procedure a large number of times M (each time randomly permuting each of the recall trials), we can obtain a distribution of control measurements $V_{\text{perm}(1)}$, $V_{\text{perm}(2)}, \ldots, V_{\text{perm}(M)}$. We can then measure the "extremity" of our empirical measurement V_{data} relative to the control distribution (e.g., by z score).

Critically, by permuting each of the recall trials, we preserve the structural quality of the empirical data (such as the recall content and the lengths of the recall trials), only allowing output order to vary. Note that across the M sets of random permutations (of each of the recall trials), any systematic output-order effects are eliminated. Thus, comparing V_{data} to the control distribution allows for isolating the focal output-order effects, free from the influence of any other biasing structural variables of the recalled data.

The rationale behind this control method is similar to that of the statistical, nonparametric permutation test (also called a *randomization test*, *rerandomization test*, or *exact test*). In the permutation test, it is possible to test whether a given experimental effect is significant with respect to a control distribution of that effect. When an effect that manifests across experimental conditions is of interest, the control distribution is generated by a repetition of the following steps: (1) randomly reshuffling (i.e., permuting) the condition labels (e.g., Condition 1, Condition 2, . . .) of the empirical observations and (2) calculating the effect of interest for the reshuffled data set, in the same manner as for the actual empirical data. Following random permutations, any difference between the conditions with respect to output position is random. Thus, the control distribution provides a "null effect" baseline, with respect to which the significance of the observed effect is gauged.

Our control procedure operates using a very similar logic. Here, the reshuffled labels are the output positions (e.g., 1, 2, 3, 4), akin to the experimental conditions in the permutation test; the empirical observations are the recalled items (designated by their serial positions at study); and the effect of interest is the RRP slope. Specifically, for each participant, we created one permutation for each of the recall trials, up to the fourth output position (because only the first four output positions participated in the present analysis) or up to the first recall error or repetition, in order to generate a "control" distribution. To illustrate, in the imaginary recall trial described above, the participant recalled five items in the order B, H, D, C, A. The first step in our analysis is to extract only Output Positions 1–4, because only these output positions participated in Farrell's analysis. Thus, we are left with B, H, D, C (with Output Position 5—A—being eliminated from the analysis). Next, we permute the output positions of these four items among themselves. Possible permutations include H, B, C, D and D, H, B, C. After each trial is randomly permuted (once), we calculated an RRP slope for the set of permuted trials—all comprising only up to the first four output positions. This calculation is identical to the calculation of the RRP slope for the actual data, with a single change: The calculations are performed on the permuted, not the actual, data. We have thus generated a single RRP slope value for our control distribution (for a single participant). By repeating this process 10,000 times, we generated 10,000 RRP slope values that constituted the null (control) distribution of RRP slopes for a single participant.

The extremity of the RRP slope per participant—controlling for the prevalence of the final list item—is given by the z score of the empirical slope relative to the control distribution. That is, the difference between the empirical slope and the mean of the control slopes is divided by the standard deviation of the control slopes. Thus, an increasing tendency for the final list item to manifest in later output positions is revealed by a positive z score, with the magnitude of the z score corresponding to the extremity of this tendency. The same is true for a decreasing tendency with a negative z score (see Howard, Youker, & Venkatadaas, 2008, for a similar approach when controlling for nonassociative tendencies that could contribute to output order). Figure 2 presents pseudocode for calculating the random-order control in the RRP analysis and for obtaining the z score for the individual participants. In the supplementary materials, we provide MATLAB code executing this calculation for the actual recall data of a single participant (number 28) from the Howard et al. (2007) data (depicted in the top panel of Fig. 3).

Fig. 2 Pseudocode for calculating the z score of an individual participant

Conduct the following for each individual participant:

1) Prepare a list of all of the recall trials for the participant. Denote the trials by $\widetilde{trial}_1, \widetilde{trial}_2, \ldots \widetilde{trial}_N$. where N is the total number of trials for the participant.
2) For each $i = 1,2,\ldots,N$:
   a) Calculate the number of output positions op_i in trial i.
   b) Find the output position e_i of the first error in trial i (the first repeated recall or recall of any word that didn't appear in the study list): If there is no error on trial i, set $e_i = \infty$
   c) Calculate the chopping output position $c_i = \min(op_i, e_i, 4)$.
   d) Chop trial i after output position c_i. Denote the chopped trial by ch_trial_i . c_i is now called the *length* of the (chopped) trial.

Comment: Following Step 2, we now have a list of error - free recall trials, each of length $c_i \leq 4$.

3) Denote $c_{max} = \max_{1 \leq i \leq N} c_i$ the 'maximal trial length'.
4) Calculate the RRP slope, rrp_slope, for the list of chopped trials by executing steps a and b below:
   a) For each $j = 1,2, \ldots c_{max}$:
      i) Calculate $actual_j$ as the number of trials in which the final list item is recalled in output position j.
      ii) Calculate $available_j$ as the number of trials with length at least j and in which the final list item is available in output position j (i.e. not recalled at prior output positions 1,2,…,j-1).
      iii) If $available_j > 0$. then calculate the j^{th} recency recall probability $RRP_j = \frac{actual_j}{available_j}$.
   b) Let l be the maximal output position j for which RRP_j was calculated.
      If $l = 1$. then the rrp slope is undefined for this participant.
      Otherwise, if $l > 1$, regress linearly (with an intercept) $RRP_1, \ldots, RRP_l$ on output positions 1, …. l. The slope of this linear regression is the RRP slope, rrp_slope.
5) For $m = 1,2, \ldots M$ (we used $M = 10{,}000$):
   a) For each $i = 1,2,\ldots,N$: Permute randomly the output positions of the elements in the chopped trial ch_trial_i: Denote the permuted trial by $trial_i^m$. Choose the permuted permutations independently across i (trial number) and m (iteration number).
   b) Calculate the simulated RRP slope, rrp_slope_m for the set of permuted trials $trial_1^m$, $trial_2^m, \ldots, trial_N^m$ by following Step 4 for the set of permuted trials instead of the actual chopped trials.
6) Calculate the mean $Mean$ and the standard deviation SD of the sequence of the simulated RRP slopes: $(rrp\_slope_1, rrp\_slope_2, \ldots, rrp\_slope_M)$.
7) Calculate the z score: $z = \frac{rrp\_slope - Mean}{SD}$.

Figure 3 illustrates the null RRP distributions of two participants from the Howard et al. (2007) study. The figure shows that the null RRP distribution for the participant in the top panel is located to the right of distribution for the participant in the bottom panel. This is caused by the fact that the nominal recency of the final list item for the "top participant" is higher. This participant recalled the final list item in 88 % of the trials, whereas the "bottom participant" recalled the final list item on a lower 42 % of the trials. This reinforces the fact that higher nominal levels of recency bias the RRP slopes to a larger extent. Next, the vertical lines correspond to the empirical RRP slopes. Both participants had positive empirical slopes: 0.3 and 0.02 for the participants in the top and bottom panels, respectively. Thus, one might have (erroneously) inferred that both participants exhibited an increasing tendency to recall the final list item as output unfolded. Critically, however, whereas the empirical slope of the "top subject" is located in the "right tail" of the null distribution—corresponding to an increased tendency to recall the final list item as output unfolds—the empirical slope for the "bottom" subject leans toward the left side of the null distribution—implying a decreased tendency. These different tendencies manifest in a positive z score (1.68), as compared to a negative z score (–0.42), for these two participants. In sum, the "signature" of a positive (uncontrolled) RRP slope can underlie both an

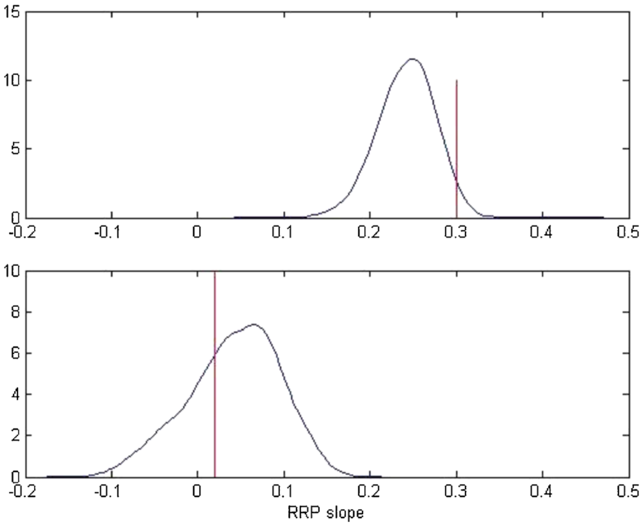

Fig. 3 Two examples of null RRP-density distributions and empirical RRP slopes (vertical lines) for two participants from the Howard et al. (2007) data set (Participants 28 and 208). The density functions were estimated with kernel estimation techniques (using the function "ksdensity" in MATLAB)

increasing and a decreasing tendency to recall the final list item later in output. Fortunately, the random-order control provides a diagnostic tool for deciphering the true direction of this tendency (see note 9 below).

Prior to analyzing the empirical data, it is important to verify that our control analysis is valid. That is, we must establish that if the output order is indeed random (hence, reflecting no tendency to recall the final list item differently across output positions), it would produce zero-mean z scores. To this end, for each value of $k = 0, 1, 2, \ldots, 20$, we simulated 10,000 participants, each participating in 20 FR trials (typical of FR studies) and recalling four outputs per trials (if the output were longer, it would in any case be truncated). The only constraint on output was that the final list item was recalled in exactly k of the 20 trials (for each k, we performed a separate simulation), and subject to this constraint, the order of the output was set to be entirely random. Thereby, we treated this simulated data as if they were "actual" data generated by participants with a constant nominal recency and random output order. Finally, we calculated for each simulated participant both an RRP slope and a control z score.

Figure 4 depicts the mean uncontrolled RRP scores across participants for each value of k. The results demonstrate our thesis that the RRP slope is strongly biased by the nominal recency for the final list item. As can be seen, although the output order is random, the mean RRP slopes monotonically increase with nominal recency and yield a value as high as .24 when the final list item is recalled on all trials. Because output order is totally random, these positive RRP slopes can be fully attributed to the bias that the nominal level of recency imposes on the uncontrolled measure of conditional recency (RRP slopes).

Critically, when we averaged the z scores across participants for these simulated data, we obtained a mean value of 0 for all values of k (the absolute mean z was smaller than 1×10^{-13} in all cases). This demonstrates that the z-score measure is effective in eliminating the bias on conditional recency that is caused by the nominal level of recency for the final list item.

An additional simulation was designed to verify that recall patterns can yield both diagnostically positive and negative control z scores. To this end, we simulated 100,000 participants, each participating in 20 recall trials and recalling four items per trial. Each participant recalled the final list item in exactly k trials [drawn uniformly from the set $(1, 2, \ldots, 20)$]. In this simulation, each time that the final list item was recalled, it appeared in the final (fourth) output position. Such a recall pattern conveys a strong tendency for the final list item to be recalled later, rather than sooner, in the output—that is, an increased tendency to be recalled as the output unfolds.

Next, for each participant we calculated both an RRP slope and a z score. The mean RRP score across all participants was strongly positive: .158. Critically, the mean z score across all participants was also highly positive—2.74—demonstrating an increasing tendency even after our random-order control. Finally, we repeated the simulation with a single change. Now, each time that the final list item was recalled, it appeared in the first (rather than the last) output position, reflecting a decreasing tendency to recall the final list item. Indeed, the mean RRP

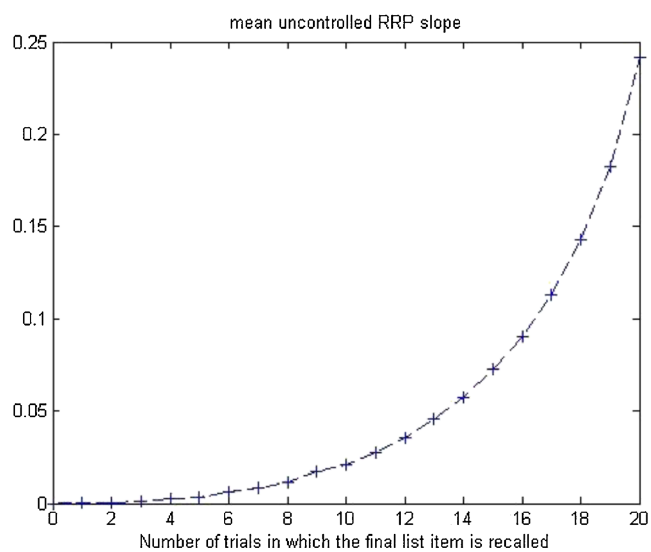

Fig. 4 Mean uncontrolled RRP slopes as a function of the number of trials (out of 20) in which the final list item was recalled in the first four output positions. Note that the output order was totally random. Therefore, these RRP slopes are entirely mediated by the bias that the nominal level of recency for the final list item imposes on the RRP slope (i.e., on conditional recency)

slope was now negative, at –.158. Critically, the mean z score remained highly negative, at –4.58. In conclusion, our control does not obscure recall tendencies to the extent of eliminating the possibility of finding inherently increasing or decreasing recall tendencies.[6]

As we explicated above, our random-order control controls for all structural components of the recall sequence, save output order. In particular, it controls not only for the prevalence of the final list item, but also for the effect of the diminishing pool of items, which was of focal concern in Farrell's (2010) control analysis. Indeed, when computing the RRP slope, the size of the available pool of items diminishes (as output unfolds) for both the empirical and the permuted data sets. Next, we will reanalyze the 14 conditions analyzed by Farrell (2010). Beyond the mathematical bias modulated by different levels of recency for the last item, might a dissociation still emerge?

Results

Figure 5 shows the mean RRP-slope z scores across participants for each condition. Conditions are grouped according to task: immediate or delayed FR. For immediate FR, only the "20–2" condition of Murdock (1962) yielded a significant positive mean z score.[7] Moreover, the "HVNK" condition (Howard et al., 2007; also immediate FR) yielded a significantly negative z score, thus testifying against an overall general positive conditional recency in immediate FR, once recall frequency of the final list item is controlled for. For the delayed-FR conditions, all of the "ISI" (Howard & Kahana, 1999, Exp. 2) conditions were significantly negative.

Next, we conducted a multilevel linear regression analysis with z scores for the individual participants as the dependent variable, task (immediate or delayed FR) as a predictor, and a random effect on the intercept for the 14 experimental conditions. The effect of immediate versus delayed recall was positive, $\beta = 0.86$. Critically, unlike Farrell's (2010) analysis, this effect did not achieve significance, $t(16.20) = 1.08$, $p = .3$. Furthermore, a paired-sample t test on the Howard and Kahana (1999, Exp. 1) data —the only experiment implementing a repeated measures design comparing immediate and delayed free recall—revealed that the mean RRP-slope z scores did not differ significantly between immediate and delayed FR (mean difference = –0.71; $t(52) = -1.37$, $p = .18$, Cohen's $d = -0.19$; note that the mean was higher for the delayed condition). In conclusion, no evidence for a dissociation between immediate and delayed FR, with respect to conditional recency, was found.[8]

Discussion

Unlike delayed FR, in immediate FR, the probability of recalling the final list item, conditional on its availability, increases as a function of output position (Farrell, 2010). Here, we controlled for the recall frequency of the final list item by reshuffling output order. We found that the dissociation between immediate and delayed FR tasks no longer emerged. Admittedly, our results with respect to the comparison between immediate and delayed FR constitute statistical null effects. Still, these null effects weaken Farrell's conclusions based on the same data set.

Unitary and dual-store theories of memory and model fits

The conditional-recency analysis is highly consequential with respect to the prolonged debate on unitary versus dual-store models of human memory (e.g., Davelaar et al., 2005; Kahana et al., 2008; Lehman & Malmberg, 2013; Sederberg et al., 2008; Usher et al., 2008). Dual-store theories posit that in addition to a long-term store, a short-term buffer, which sustains the (otherwise decaying) activation of the recently encountered items, is operative in immediate FR. Farrell (2010) demonstrated that the increasing conditional recency in immediate FR challenges unitary models of episodic memory

[6] In these last two simulations (i.e., when the final list item was recalled, it was always recalled in the last output position or in the first output position), we had to adjust the values of k. In both simulations, we did not use the value $k = 0$. This was because when the final list item is never recalled, the z score is not defined. [For any permutation of the trials, each recency recall probability $RRP(j) = 0$. Thus, the RRP slope is always 0, and the standard deviation of the null RRP distribution is also 0, leaving the z score undefined.] In addition, in the simulation in which the final list item was always recalled in Output Position 1, we dropped the value of 20, because for this value the final list item was never available following Output Position 1, and therefore the RRP slope also could not be calculated (see Step 4b in the pseudocode in Fig. 2).

[7] When a Bonferroni correction for multiple comparisons was applied, this effect no longer achieved significance.

[8] For some of the participants, the null RRP distributions were moderately skewed (see Fig. 3). To exclude the possibility that this biased our conclusions, we repeated all of our analyses, but instead of using the z score, the dependent measure was obtained by only subtracting the mean of the null RRP distribution from the empirical RRP slope (but not dividing by its standard deviation). Hence, in Step 7 of the pseudocode, we calculated only the numerator, and not the denominator. When using this measure, all of our results were similar, with no change whatsoever to the conclusion. For instance, in the multilevel linear regression analysis, the effect of immediate versus delayed recall was positive ($\beta = 0.04$) but not significant, $t(18.95) = 1.40$, n.s. Additionally, the paired-sample t test comparing the immediate-FR and delayed-FR conditions of Experiment 1 of Howard and Kahana (1999) revealed no significant difference [difference = –0.01; $t(55) = -0.17$, n.s.]. Finally, note that in our first simulation with a random output order, the mean of the z scores was 0 for all levels of recency, implying that the z-score measure did not introduce a systematic bias into our analysis.

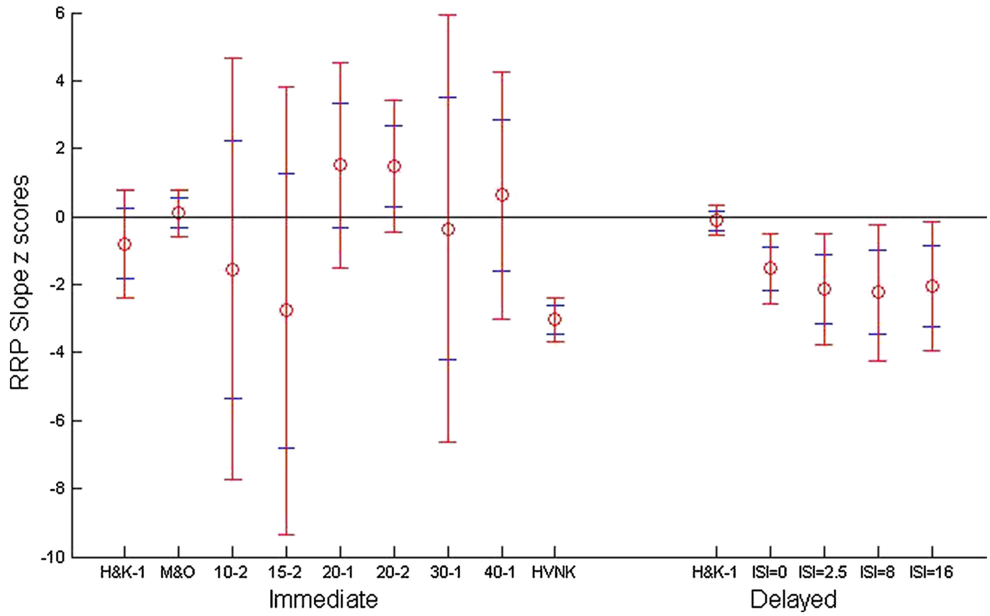

Fig. 5 Mean RRP-slope z scores for Output Positions 1–4. The arrangement and labeling of the conditions are identical to those in Fig. 1. The inner error bars show single-sample 95 % confidence intervals, whereas the outer error bars shows 95 % confidence intervals with Bonferroni correction for multiple comparisons

such as the TCM and SIMPLE, in that computational fits of these models to empirical data predict negative conditional-recency slopes. This occurs because in these models, the passage of time during the recall phase diminishes either the overlap between the encoding context and the evolving test context of the final list item (relative to prior items; TCM) or the enhanced discriminability of the encoding context (SIMPLE) (see note 2). In addition, Farrell demonstrated that computational fits of the model to the empirical data were improved when assorted mechanisms of STM were incorporated into the models.

The inability of the TCM to account satisfactorily for the conditional-recency dissociation (and the better accounts provided by an STM mechanism) has been argued to indicate that dual-store models are superior to unitary models (Usher et al., 2008; see note 1). Nevertheless, our findings suggest that the conditional-recency dissociation is too compromised to serve as a deciding factor when assessing unitary against dual-store models of memory.

Implications with respect to additional analyses described by Farrell (2010)

Farrell (2010) conducted several analyses that went beyond conditional-recency slopes. In these, he showed that rather than only looking at the slope of the RRP function, it is useful to look at the shape of the function (the "RRP profile"). He argued that different experimental conditions create meaningfully different profiles. Specifically, in some immediate-FR conditions, RRP curves initially increase and later decrease, whereas in other conditions the opposite occurs, with an early decrease (or flatness) followed by an increase (see Farrell, 2010, Fig. 2). Additionally, Farrell demonstrated how these RRP profiles could be used to challenge FR models for adequate fit, and thus could serve the purpose of model diagnosis.

Farrell (2010) used the RRP functions to compare alternative candidate mechanisms for recall buffers, if such recall buffers are indeed involved in FR tasks (Sederberg et al., 2008; Usher et al., 2008). Importantly, by utilizing a model comparison approach, Farrell demonstrated that models that incorporated certain STM mechanisms (e.g., a "forward buffer") outperformed purely unitary models in most of the immediate-FR conditions (and were outperformed by TCM in most of the delayed-FR conditions). With respect to these additional sets of analysis, we wish to make the following comments.

First, the functional findings (the RRP profiles) did not yield an empirical dissociation between immediate and delayed FR. Moreover, these findings exhibited substantial differences within a task (e.g., immediate FR). We thus find the conditional-slope dissociation to be the most dramatic and coherent empirical finding of Farrell's (2010) analysis, the interpretation of which has now been put in question.

Second, as we have explicitly argued in this article, the random-order control method is not limited to RRP slopes. Rather, it can be applied to any analysis that makes claims regarding the theoretical tendencies that manifest across output positions. Because RRP functions are a case in point, future RRP function analyses might benefit from such a control.

We now consider the model comparison approach that Farrell (2010) conducted to compare unitary and dual-store models. We find this approach perfectly valid and convincing.[9] Importantly, such model comparisons are based on the entire recall sequences and are not limited to conditional-recency patterns (slopes or functions). They thus provide general findings that are of intrinsic value, beyond any aspects of conditional recency. In this respect, Farrell's finding that incorporating STM buffers into unitary models such as TCM enhances their ability to account for recall data is highly important and indicative for the unitary- versus dual-store debate.

Still, although it is indicative, we encourage a cautious and preliminary interpretation of these findings. Farrell's analysis is restricted to the "standard TCM" model. A more sophisticated and flexible version of TCM, albeit one that is more complex with respect to parameter economics—the TCM-A—has been presented (Sederberg et al., 2008). It is still unknown how this new model will fares with respect to conditional-recency data. Also, as was noted by Farrell (2010), the unitary models with a buffer supplement that he evaluated "do not constitute full models of free recall" (p. 342). So, conclusions with respect to these model comparisons cannot yet be definitively made.

A confound or a useful redescription?

In this article, we showed that the RRP slope is highly modulated by the prevalence of the final list item. Yet, some scholars may note that it is often the case that co-varying measures are successfully and simultaneously used in the study of a given phenomenon (e.g., in FR, both serial position and the probability of first recall are often used). Thus, it may be argued that the (uncontrolled) RRP slope provides useful descriptions of empirical data, while acknowledging that part of this description overlaps with nominal recency.

Here, however, we have made the case that the tendency to recall the final list item across output positions is biased, and hence should better be isolated from the prevalence of the final list item. Importantly, our work should be of theoretical interest even for those who interpret what we find to be a measurement bias as a valid part of their construct. Indeed, a comparison between Farrell's (2010) original analysis and our reanalysis of the same data sets implicates the extent to which the conditional-recency dissociation is affected by the nominal recency of the final list item. As such, this comparison should be useful for students of conditional recency in furthering the understanding of its underlying mechanisms.

[9] Such an approach was implemented by Farrell (2010) by maximum likelihood estimation of free parameters for the alternative models, followed by calculation of a "model-complexity compensating" criterion, such as the Bayesian information criterion, and selecting the model that achieved the minimum criterion among the model candidates.

Author Note We thank Simon Farrell for his conscientious comments on an earlier draft of this article and for his help in thinking through the data during the revision process. We also thank Marius Usher, Tomer Carmel, Andrei Teodorescu, and Zohar Bronfman for providing helpful comments about the manuscript.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*, 173–189. doi:10.1016/0010-0285(74)90009-7

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576. doi:10.1037/0033-295X.114.3.539

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*, 3–42. doi:10.1037/0033-295X.112.1.3

Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall: A challenge for unitary models of recency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 324–347. doi:10.1037/a0018042

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360. doi:10.1016/S0022-5371(66)80044-0

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., Gretz, A. L., . . . Turpin, B. M. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 355–369. doi:10.1037/0278-7393.6.4.355

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 923–941. doi:10.1037/0278-7393.25.4.923

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299. doi:10.1006/jmps.2001.1388

Howard, M. W., Venkatadass, V., Norman, K. A., & Kahana, M. J. (2007). Associative processes in immediate recency. *Memory & Cognition*, *35*, 1700–1711. doi:10.3758/BF03193503

Howard, M. W., Youker, T. E., & Venkatadass, V. S. (2008). The persistence of memory: Contiguity effects across hundreds of seconds. *Psychonomic Bulletin & Review*, *15*, 58–63. doi:10.3758/PBR.15.1.58

Kahana, M. J., Sederberg, P. B., & Howard, M. W. (2008). Putting short-term memory into context: Reply to Usher, Davelaar, Haarmann, and Goshen-Gottstein (2008). *Psychological Review*, *115*, 1119–1126. doi:10.1037/a0013724

Lehman, M., & Malmberg, K. J. (2013). A buffer model of encoding and temporal correlations in retrieval. *Psychological Review*, *120*, 155–189. doi:10.1037/a0030851

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488. doi:10.1037/h0045106

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, *86*, 263–267. doi:10.1037/h0029993

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, *17*, 132–138. doi:10.1080/17470216508416422

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*, 893–912. doi:10.1037/a0013396

Usher, M., Davelaar, E. J., Haarmann, H. J., & Goshen-Gottstein, Y. (2008). Short-term memory after all: Comment on Sederberg, Howard, and Kahana (2008). *Psychological Review*, *115*, 1108–1118. doi:10.1037/a0013725